# Foundation metrics for evaluating effectiveness of healthcare conversations powered by generative AI

Year: 2023 | Citations: 116 | Authors: Mahyar Abbasian, Elahe Khatibi, Iman Azimi, David Oniani, Zahra Shakeri Hossein Abad

## Abstract

Generative Artificial Intelligence is set to revolutionize healthcare delivery by transforming traditional patient care into a more personalized, efficient, and proactive process. Chatbots, serving as interactive conversational models, will probably drive this patient-centered transformation in healthcare. Through the provision of various services, including diagnosis, personalized lifestyle recommendations, dynamic scheduling of follow-ups, and mental health support, the objective is to substantially augment patient health outcomes, all the while mitigating the workload burden on healthcare providers. The life-critical nature of healthcare applications necessitates establishing a unified and comprehensive set of evaluation metrics for conversational models. Existing evaluation metrics proposed for various generic large language models (LLMs) demonstrate a lack of comprehension regarding medical and health concepts and their significance in promoting patients' well-being. Moreover, these metrics neglect pivotal user-centered aspects, including trust-building, ethics, personalization, empathy, user comprehension, and emotional support. The purpose of this paper is to explore state-of-the-art LLM-based evaluation metrics that are specifically applicable to the assessment of interactive conversational models in healthcare. Subsequently, we present a comprehensive set of evaluation metrics designed to thoroughly assess the performance of healthcare chatbots from an end-user perspective. These metrics encompass an evaluation of language processing abilities, impact on real-world clinical tasks, and effectiveness in user-interactive conversations. Finally, we engage in a discussion concerning the challenges associated with defining and implementing these metrics, with particular emphasis on confounding factors such as the target audience, evaluation methods, and prompt techniques involved in the evaluation process.